Комбинаторные оценки вероятности переобучения на основе кластеризации и покрытий множества алгоритмов*

Фрей А. И., Толстихин И. О.
sashafrey@gmail.com, iliya.tolstikhin@gmail.com
Вычислительный Центр им. А. А. Дородницына РАН

В данной работе предлагается новая комбинаторная оценка вероятности переобучения, учитывающая сходство алгоритмов. Оценка основана на разложении множества алгоритмов на непересекающиеся подмножества (кластеры). Итоговая оценка учитывает сходство алгоритмов внутри каждого кластера, и расслоение алгоритмов по числу ошибок между кластерами. Для оценки вероятности переобучения каждого кластера предлагается теоретико-групповой подход, основанный на учете симметрий. На примере задач из репозитория UCI показано, что предлагаемый метод в ряде случаев дает менее завышенную оценку вероятности переобучения по сравнению с известными ранее комбинаторными оценками.

Combinatorial bounds on probability of overfitting based on clustering and coverage of classifiers.*

Frey A. I., Tolstikhin I. O.
Computing Centre of RAS

The paper improves existing combinatorial bounds on probability of overfitting. A new bound is based on partitioning of a set of classifiers into non-overlapping clusters, and then embedding each cluster into a superset with known exact formula for the probability of overfitting. The key idea is to account for similarities between classifiers within each cluster. As a result, the new bound outperforms existing combinatorial bounds in our experiments on real datasets from UCI repository.

Введение

Решение задач классификации и прогнозирования можно рассматривать как задачу выбора по неполной информации. Качество алгоритма, выбранного по конечной обучающей выборке объектов, часто оказывается значительно хуже на независимой контрольной выборке. В таких случаях говорят, что произошло *переобучение* алгоритма [1, 2].

В комбинаторной теории оценок обобщающей способности [3] *вероятностью переобучения* называют долю разбиений генеральной выборки на обучающую и контрольную подвыборки фиксированной длины, при которых произошло переобучение. В [4] показано, что вероятность переобучения зависит от *профиля расслоения* множества алгоритмов и от *сходства алгоритмов* между собой.

Профилем расслоения называют распределение алгоритмов по числу ошибок на генеральной выборке. Алгоритмы с высоким числом ошибок имеют низкую вероятность реализоваться в результате обучения и потому дают незначительный вклад в вероятность переобучения. Данное явление подробно изучено в работах [5, 6, 7]. В дальнейших работах [8, 9] показано, что применение полученных оценок вероятности переобучения позволяет

Работа поддержана РФФИ (проект № 11-07-00480, № 12-07-33099-мол-а-вед) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

значительно улучшить качество композиций логических алгоритмов классификации на многих задачах из репозитория UCI.

Схожими называют алгоримы, хэмминговы расстояния между которыми малы. В [4] экспериментально показано, что сходство алгоритмов существенно уменьшает вероятность переобучения. Отметим, что большинство комбинаторных оценок вероятности переобучения основаны на анализе пар связных алгоритмов, т.е. различающихся только на одном объекте. Данный подход не позволяет в полной мере учесть сходство алгоритмов. Предложенный в [10] метод позволяет учесть сходство, но лишь в виде оценок худшего случая.

В данной работе связь между сходством алгоритмов и вероятностью переобучения будет изучена с помощью теоретико-группового подхода [11, 12, 13]. Чтобы устранить эффект расслоения в первую очередь будут рассмотрены модельные семейства, в которых все алгоритмы допускают равное число ошибок на полной выборке. На основе полученных результатов предлагается новая верхняя оценка вероятности переобучения, основанная на кластеризации алгоритмов с близкими векторами ошибок. Затем данная оценка обобщается на семейства с расслоением алгоритмов по числу ошибок. Эксперименты на 11 задачах из репозитория UCI показывают, что предлагаемый подход в ряде случаев уточняет все известные оценки вероятности переобучения.

Определения

Пусть задана генеральная выборка $\mathbb{X} = \big(x_1, \ldots, x_L\big)$, состоящая из L объектов. Произвольный алгоритм классификации, примененный к данной выборке, порождает бинарный вектор ошибок $a \equiv \big(I(a, x_i)\big)_{i=1}^{L}$, где $I(a, x_i) \in \{0, 1\}$ — бинарный индикатор ошибки алгоритма a на объекте x_i. В дальнейшем генеральная выборка $\mathbb{X}$ предполагается фиксированной, поэтому алгоритмы будут отождествляться с векторами их ошибок на выборке $\mathbb{X}$.

Для произвольной подвыборки $U \subseteq \mathbb{X}$ *число* и *частота* ошибок алгоритма a обозначаются, соответственно, через $n(a, U) = \sum\limits_{x_i \in U} I(a, x_i)$ и $\nu(a, U) = n(a, U)/|U|$.

Пусть $[\mathbb{X}]^\ell$ — множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$. *Методом обучения* называют отображение μ, которое произвольной обучающей выборке $X \in [\mathbb{X}]^\ell$ ставит в соответствие некоторый алгоритм $a = \mu(A, X)$ из заранее фиксированного множества $A \subset \mathbb{A}$, где $\mathbb{A} = \{0, 1\}^L$ — множество всех возможных бинарных векторов ошибок. Для произвольного разбиения $X \sqcup \bar{X} = \mathbb{X}$ *переобученностью* алгоритма $a = \mu(A, X)$ называют уклонение частот его ошибок на контроле и на обучении $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

Частоту ошибок $\nu(a, X)$ алгоритма a на обучающей выборке X часто называют *эмпирическим риском*. *Минимизация эмпирического риска* (МЭР) — это такой метод обучения, что для любой обучающей выборки X выбранный алгоритм $a = \mu(A, X)$ допускает наименьшее число ошибок на обучающей выборке X. Таким образом, для всех $X \in [\mathbb{X}]^\ell$ должно быть выполнено $\mu(A, X) \in A(X)$, где

$$A(X) \equiv \operatorname{Arg}\min_{a \in A} n(a, X). \tag{1}$$

При минимизации эмпирического риска может возникать неоднозначность — несколько алгоритмов из $A(X)$ могут иметь одинаковое число ошибок на обучающей выборке. Для устранения неоднозначности используется метод *пессимистической минимизации эмпирического риска*, выбирающий в $A(X)$ алгоритм с наибольшим числом ошибок на полной выборке. Пессимистический МЭР не может быть реализован на практике, т.к. он подглядывает в скрытую часть генеральной выборки. Тем не менее, пессимистический МЭР

является удобной теоретической конструкцией, поскольку он позволяет получать верхние оценки на вероятность переобучения любого МЭР.

Следуя слабой вероятностной аксиоматике [4], будем считать, что на множестве $[\mathbb{X}]^\ell$ всех C_L^ℓ разбиений $X \sqcup \bar{X}$ введено равномерное распределение вероятностей. Тогда вероятность переобучения $Q_\varepsilon(A)$ определяется как доля разбиений, при которых переобученность превышает заданный порог $\varepsilon \in (0,1]$:

$$Q_\varepsilon(A) = \mathsf{P}[\delta(\mu(A,X),X) \geqslant \varepsilon], \tag{2}$$

где $\mathsf{P}[\varphi] \equiv \frac{1}{C_L^\ell} \sum\limits_{X \in [\mathbb{X}]^\ell} \varphi(X)$, φ — произвольный предикат на множестве разбиений $[\mathbb{X}]^\ell$, а квадратные скобки переводят логическое выражение в число 0 или 1 по правилам [истина] $= 1$, [ложь] $= 0$. Заметим, что $1 - Q_\varepsilon(A)$ как функция порога ε есть функция распределения случайной величины $\delta(\mu(A,X),X)$, определенной на конечном вероятностном пространстве $\{[\mathbb{X}]^\ell, 2^{[\mathbb{X}]^\ell}, \mathsf{P}\}$, где P — равномерное распределение. В случаях, когда из контекста понятно, о каком множестве алгоритмов идёт речь, будем опускать аргумент A и записывать вероятность переобучения как Q_ε.

Рассмотрим множество $A = \{a\}$, состоящее из одного алгоритма. Тогда $\mu X = a$ для любой выборки $X \in [\mathbb{X}]^\ell$. Это значит, что вероятность переобучения Q_ε преобразовалась в вероятность больших уклонений между частотами ошибок в выборках $X, \bar{X}$. Допустив, что число ошибок $n(a,\mathbb{X})$ нам известно, получим точное выражение для Q_ε.

Теорема 1 (FC-оценка [8]). *Для фиксированного алгоритма a, такого что $m = n(a,\mathbb{X})$, и любого $\varepsilon \in [0,1]$ вероятность переобучения определяется левым хвостом гипергеометрического распределения:*

$$Q_\varepsilon(a) = H_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right), \tag{3}$$

где $H_L^{\ell,m}(s) = \sum\limits_{t=0}^{s} h_L^{\ell,m}(t)$ — функция гипергеометрического распределения, $h_L^{\ell,m}(t) = C_m^t C_{L-m}^{\ell-t}/C_L^\ell$ — функция плотности гипергеометрического распределения [4].

Гипергеометрическое распределение играет важную роль во многих комбинаторных оценках. Оценка (3), примененная совместно с неравенством Буля, позволяет получить верхнюю оценку на $Q_\varepsilon(A)$, справедливую для любого метода обучения μ.

Теорема 2 (VC-оценка [8]). *Для любого метода обучения μ и любого $\varepsilon \in [0,1]$ вероятность переобучения ограничена суммой FC-оценок по множеству алгоритмов A:*

$$Q_\varepsilon(A) \leqslant \mathsf{P}\big[\max_{a \in A} \delta(a,X) \geqslant \varepsilon\big] \leqslant \sum_{a \in A} H_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right), \quad m = n(a,\mathbb{X}). \tag{4}$$

Назовём две причины завышенности оценки (4). Во-первых, большинство алгоритмов из A имеют высокую частоту ошибок и, следовательно, имеют исчезающе малую вероятность реализоваться в результате обучения. Тем не менее, оценка равномерного уклонения игнорирует это свойство метода обучения μ. Во-вторых, неравенство Буля игнорирует тот факт, что алгоритмы с близкими векторами ошибок переобучаются в основном на одних и тех же разбиениях. Более точные оценки должны учитывать свойства метода обучения и сходство между алгоритмами.

Эффект сходства

Хэмминговым расстоянием между алгоритмами a_1 и a_2 называют величину

$$\rho(a_1,a_2) = \sum_{x_i \in \mathbb{X}} [a_1(x_i) \neq a_2(x_i)].$$

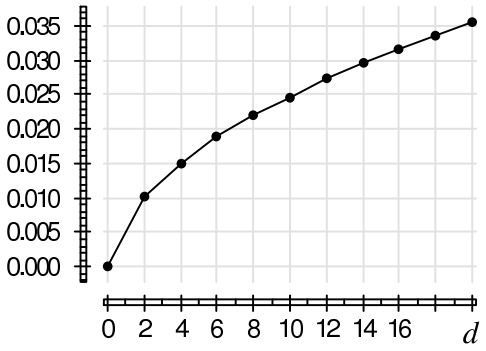

Рис. 1. Зависимость медианы распределения Q_ε от хэммингова расстояния $d = \rho(a_1, a_2)$ между векторами ошибок пары алгоритмов. $L = 100$, $\ell = 50$.

Рассмотрим простой пример, иллюстрирующий зависимость переобучения от хэммингова расстояния между алгоритмами семейства.

Эксперимент 1. Множество $A = (a_1, a_2)$ состоит из двух алгоритмов, допускающих по m ошибок на полной выборке. Векторы ошибок подобраны так, чтобы хэммингово расстояние $\rho(a_1, a_2)$ равнялось заранее фиксированному числу d. На рисунке 1 приведена зависимость медианы распределения $Q_\varepsilon(A)$ от хэммингова расстояния между алгоритмами. Видно, что переобучение увеличивается с ростом $\rho(a_1, a_2)$.

Зададимся целью построить верхнюю оценку вероятности переобучения, учитывающую данный эффект. Допустим, что исходное множество алгоритмов A представлено в виде разбиения на непересекающиеся подмножества $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ так, что в каждое A_i попали лишь алгоритмы с близкими векторами ошибок. В данной ситуации будем называть множества A_i *кластерами* алгоритмов. Покажем, что задачу оценивания вероятности переобучения всего множества A можно свести к оцениванию вероятности переобучения отдельных кластеров.

Лемма 3. *Пусть множество алгоритмов A произвольным образом представлено в виде разбиения на непересекающиеся подмножества $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$. Тогда вероятность переобучения пессимистического метода минимизации эмпирического риска оценивается сверху следующим выражением:*

$$Q_\varepsilon(A) \leqslant \sum_{i=1}^{t} Q_\varepsilon(A_t). \tag{5}$$

В дальнейшем лемма 3 будет играть ту же роль, что и неравенство Буля при выводе оценки (4). Преимущество данной леммы в том, что вместо суммирования по всем алгоритмам $a \in A$ суммирование производится по кластерам произвольного разбиения $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$.

Доказательство. Заметим, что достаточно доказать неравенство (5) для $t = 2$ (для произвольного числа кластеров неравенство доказывается индукцией по t). Обозначим через $\mu(A, X)$ алгоритм, выбранный пессимистическим методом минимизации эмпирического риска из множества A по обучающей выборке X. Рассмотрим произвольное разбиение $X \in [\mathbb{X}]^\ell$ и покажем следующее:

$$[\delta(\mu(A, X), X) \geqslant \varepsilon] \leqslant [\delta(\mu(A_1, X), X) \geqslant \varepsilon] + [\delta(\mu(A_2, X), X) \geqslant \varepsilon]. \tag{6}$$

Для разбиения X и множеств A_1, A_2, A множества $A_1(X), A_2(X), A(X)$ определены согласно (1). Обозначим через $n_1(X), n_2(X)$ и $n(X)$ число ошибок на обучающей выборке для алгоритмов из $A_1(X)$, $A_2(X)$ и $A(X)$, соответственно. Очевидно, что $n_1(X) \geqslant n(X)$ и $n_2(X) \geqslant n(X)$, но по крайней мере одно из этих неравенств обязательно обращается в равенство. Рассмотрим два случая: в первом одно неравенство строгое, во втором оба неравенства обращаются в равенство.

Случай 1. Пусть для определенности $n_1(X) > n(X)$. Тогда $A_2(X) = A(X)$, и следовательно $\mu(A_2, X) = \mu(A, X)$, откуда немедленно следует (6).

Случай 2. Из $n_1(X) = n_2(X) = n(X)$ следует, что $A(X) = A_1(X) \cup A_2(X)$, и, таким образом, либо $\mu(A, X) \in A_1(X)$, либо $\mu(A, X) \in A_2(X)$ (в зависимости от того, в какое из этих двух множеств попал алгоритм с наибольшим числом ошибок на полной выборке). Значит, вновь выполнено (6). ■

Для вывода верхних оценок вероятности переобучения каждого кластера удобно потребовать, чтобы внутри каждого кластера алгоритмы допускали равное число ошибок на полной выборке. Тогда можно воспользоваться следующей леммой из работы [13].

Лемма 4. *Пусть A_i, B — два множества алгоритмов, $A_i \subseteq B$, и все алгоритмы из B допускают равное число ошибок на полной выборке. Пусть метод обучения является минимизацией эмпирического риска. Тогда для всех $\varepsilon > 0$ выполнено неравенство $Q_\varepsilon(A_i) \leqslant Q_\varepsilon(B)$.*

Доказательство. Докажем утверждение для частного случая $B = A_i \cup \{b\}$. Рассмотрим произвольное разбиение $X \in [\mathbb{X}]^\ell$. Нас интересуют только разбиения с $\mu(B, X) = b$, потому что вклад остальных разбиений в вероятность переобучения не изменился. Пусть $a = \mu(A_i, X)$ — алгоритм, выбранный на разбиении X методом обучения из множества A_i. Поскольку μ является минимизацией эмпирического риска, получим $n(b, X) \leqslant n(a, X)$. Поскольку по условию алгоритмы a и b имеют равное число ошибок на полной выборке, уклонение частоты $\delta(b, X) \geqslant \delta(a, X)$. Следовательно, вклад каждого разбиения от добавления алгоритма b мог только увеличиться. ■

На выбор множества B накладывается несколько условий. Во-первых, оно должно состоять из алгоритмов с равным числом ошибок на полной выборке. Во-вторых, все алгоритмы должны иметь близкие векторы ошибок. В-третьих, оценка вероятности переобучения $Q_\varepsilon(B)$ должна быть вычислительно эффективной. Оказывается, следующее семейство удовлетворяет всем этим требованиям.

Определение 1. *Пусть a_0 — произвольный алгоритм с m ошибками, $r \leqslant m$ — натуральное число. Центральным слоем хэммингова шара называется множество:*

$$B_r^m(a_0) = \{a \in \mathbb{A} \colon \rho(a, a_0) \leqslant r \text{ и } n(a, \mathbb{X}) = m\}.$$

Данное множество состоит из алгоритмов хэммингова шара радиуса r с центром в a_0, допускающих на полной выборке столько же ошибок, сколько и центр шара. Вероятность переобучения $Q_\varepsilon(B_r^m(a_0))$ зависит только от радиуса шара r и числа ошибок $n(a_0, \mathbb{X})$, поэтому в дальнейшем вместо $B_r^m(a_0)$ будет использоваться сокращенная запись B_r^m.

Эксперимент 2. Исследуем вероятность переобучения $Q_\varepsilon(B_r^m)$ численно с помощью метода Монте-Карло, сэмплируя (2) по 10 тыс. случайным подвыборкам $X \in [\mathbb{X}]^\ell$. Для сравнения мы рассмотрим еще одно модельное множество R_n^m, составленное из n алгоритмов, допускающих по m ошибок на полной выборке. Векторы ошибок всех алгоритмов из R_n^m сгенерированы случайно и независимо. В следующей таблице показано, при каком числе алгоритмов n в R_n^m медианы распределений $Q_\varepsilon(R_n^m)$ и $Q_\varepsilon(B_r^m)$ совпадают.

Таблица 1. Сравнение $|R_n^m|$ и $|B_r^m|$ при $L = 50$, $\ell = 25$, $m = 10$.

r	$\lvert B_r^m\rvert$	$\lvert R_n^m\rvert$	δ
2	401	2	0.079
4	35 501	7	0.160
6	1 221 101	39	0.240
8	20 413 001	378	0.319

Из таблицы 1 видно, что всего семь алгоритмов со случайными векторами ошибок могут переобучиться так же сильно, как и множество из десятков тысяч алгоритмов с близкими векторами ошибок. В следующем параграфе мы детально изучим это свойство множества B_r^m и приведем точную формулу для вероятности переобучения $Q_\varepsilon(B_r^m)$.

Центральный слой хэммингова шара

Точная формула вероятности переобучения $Q_\varepsilon(B_r^m)$ впервые приводится в работе [13].

Теорема 5 ([13]). *Вероятность переобучения рандомизированного метода минимизации эмпирического риска для m-го слоя хэммингова шара B_r^m радиуса r при $r \leqslant 2m$ и $n(a_0, \mathbb{X}) = m$ записывается в виде*

$$Q_\varepsilon(B_r^m) = H_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k) + \left\lfloor r/2 \right\rfloor\right) \cdot [m \geqslant \varepsilon k], \tag{7}$$

где $H_L^{\ell,m}(s)$ — функция гипергеометрического распределения.

Доказательство этого утверждения ранее не публиковалось, и мы приводим его в настоящем параграфе. При доказательстве будет использоваться теоретико-групповой подход [11, 12, 13].

Пусть $S_L = \{\pi\colon \mathbb{X} \to \mathbb{X}\}$ — симметрическая группа из L элементов, действующая на генеральную выборку перестановками объектов. Действие произвольной $\pi \in S_L$ на алгоритм $a \in A$ определено перестановкой координат вектора ошибок: $(\pi a)(x_i) = a(\pi^{-1}x_i)$. Для произвольной выборки $X \in [\mathbb{X}]^\ell$ и множества алгоритмов $A \subset \{0,1\}^L$ действия πX и πA определены следующим образом: $\pi X = \{\pi x\colon x \in X\}$, $\pi A = \{\pi a\colon a \in A\}$.

Определение 2 ([11]). *Группой симметрий* $\mathrm{Sym}(A)$ *множества алгоритмов $A \subset \mathbb{A}$ называют его стационарную подгруппу:*

$$\mathrm{Sym}(A) = \{\pi \in S_L\colon \pi A = A\}.$$

Орбитой элемента m множества M, на котором действует группа G, называется подмножество $Gm = \{gm\colon g \in G\} \subseteq M$. Орбиты двух элементов m_1 и m_2 либо не пересекаются, либо совпадают. Это позволяет говорить о разбиении множества M на непересекающиеся орбиты: $M = Gm_1 \sqcup \ldots \sqcup Gm_k$. Множество орбит действия группы $\mathrm{Sym}(A)$ на A обозначим через $\Omega(A)$, и для каждой орбиты $\omega \in \Omega(A)$ обозначим через a_ω произвольного представителя этой орбиты. Аналогично, множество орбит действия $\mathrm{Sym}(A)$ на $[\mathbb{X}]^\ell$ обозначим через $\Omega([\mathbb{X}]^\ell)$, и для орбиты $\tau \in \Omega([\mathbb{X}]^\ell)$ обозначим её представителя через X_τ.

Для широкого класса методов обучения группа $\mathrm{Sym}(A)$ позволяет учесть симметрии множества алгоритмов при выводе оценок вероятности переобучения. В частности, в работах [11, 12] используется *рандомизированный метод минимизации эмпирического риска*.

Этот метод выбирает произвольный алгоритм из множества $A(X)$ случайно и равновероятно. Определение вероятности переобучения (2) приходится модифицировать, усреднив переобученность по множеству $A(X)$:

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \frac{1}{|A(X)|} \sum_{a \in A(X)} \big[\delta(a, X) \geqslant \varepsilon\big]. \tag{8}$$

В [12] показано, что с учетом группы симметрий $\mathrm{Sym}(A)$ вероятность переобучения (8) может быть переписана следующим образом:

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{\omega \in \Omega(A)} |\omega| \sum_{X \in [\mathbb{X}]^\ell} \frac{[a_\omega \in A(X)]}{|A(X)|} [\delta(a_\omega, X) \geqslant \varepsilon].$$

Для вывода точной формулы $Q_\varepsilon(B_r^m)$ нам будет удобнее рассматривать действие группы симметрий на множестве $[\mathbb{X}]^\ell$ всех разбиений выборки на обучение и контроль.

Лемма 6. *Пусть для некоторой функции $f\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \mathbb{R}$ для всех $A \subset \mathbb{A}$, $X \in [\mathbb{X}]^\ell$ и всех $\pi \in \mathrm{Sym}(A)$ выполнено условие $f(A, X) = f(A, \pi X)$. Тогда справедливо следующее разложение:*

$$\sum_{X \in [\mathbb{X}]^\ell} f(A, X) = \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| f(A, X_\tau).$$

Доказательство. Доказательство с очевидностью следует из группировки равных слагаемых. ∎

Нас интересует, какие функции удовлетворяют условию $f(A, X) = f(A, \pi X)$. Введем следующую классификацию:

- Симметричной функцией *первого рода* будем называть $g\colon \mathbb{A} \times [\mathbb{X}]^\ell \to \mathbb{R}$, такую что для всех $\pi \in S_L$ выполнено $g(a, X) = g(\pi a, \pi X)$;
- Симметричной функцией *второго рода* будем называть $G\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to 2^{\mathbb{A}}$, такую что для всех $\pi \in S_L$ выполнено $\pi G(A, X) = G(\pi A, \pi X)$;
- Симметричной функцией *третьего рода* будем называть $f\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \mathbb{R}$, такую что для всех $\pi \in S_L$ выполнено $f(A, X) = f(\pi A, \pi X)$.

В [12] было показано, что функции $n(a, X)$ и $\nu(a, X)$ являются симметричными функциями первого рода, а A и $A(X)$ — симметричными второго рода. Следующие две теоремы позволяют легко строить новые симметричные функции из уже имеющихся.

Теорема 7. *Пусть $g_1, g_2, \ldots, g_p$ — симметричные функции первого рода, $f_1, f_2, \ldots, f_p$ — симметричные функции третьего рода, $F\colon \mathbb{R}^p \to \mathbb{R}$ — произвольная функция многих переменных. Тогда $F(g_1, g_2, \ldots, g_p)$ — вновь симметричная функция первого рода, $F(f_1, f_2, \ldots, f_p)$ — симметричная функция третьего рода.*

Доказательство. Проведя элементарные выкладки, получим

$$F(\pi a, \pi X) \equiv F(g_1(\pi a, \pi X), \ldots, g_p(\pi a, \pi X)) = F(g_1(a, X), \ldots, g_p(a, X)) \equiv F(a, X),$$

и аналогично для функций третьего рода. ∎

Теорема 8. *Пусть g — симметричная функция первого рода, G — симметричная функция второго рода. Тогда $f(A, X) \equiv |G(A, X)|$ и $f(A, X) \equiv \sum\limits_{a \in G(A,X)} g(a, X)$ — симметричные функции третьего рода.*

Доказательство. Заметим, что для любого $A \subset \mathbb{A}$ выполнено $|A| = |\pi A|$, поскольку π, как элемент группы, является биекцией. Следовательно, $|G(A,X)| = |\pi G(A,X)| = |G(\pi A, \pi X)|$.

Для функции $f(A,X) \equiv \sum\limits_{a \in G(A,X)} g(a,X)$ запишем следующую цепочку равенств:

$$f(\pi A, \pi X) = \sum_{a \in G(\pi A, \pi X)} g(a, \pi X) = \sum_{a \in \pi G(A,X)} g(a, \pi X) =$$
$$= \sum_{a \in G(A,X)} g(\pi a, \pi X) = \sum_{a \in G(A,X)} g(a,X) = f(A,X).$$

■

Из приведенных выше теорем следует, что $\frac{1}{|A(X)|} \sum\limits_{a \in A(X)} [\delta(a,X) \geqslant \varepsilon]$ является симметричной функцией третьего рода, а следовательно вероятность переобучения можно факторизовать по действию группы симметрий на множестве разбиений выборки:

$$Q_\varepsilon(A) = \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} \frac{|\tau|}{|A(X_\tau)|} \sum_{a \in A(X_\tau)} [\delta(a, X_\tau) \geqslant \varepsilon]. \tag{9}$$

Данную формулу можно упростить для случая, когда все алгоритмы из A имеют равное число ошибок на полной выборке.

Теорема 9. *Пусть все $a \in A$ имеют равное число ошибок на полной выборке. Тогда вероятность переобучения рандомизированного метода минимизации эмпирического риска записывается в виде*

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \left[\min_{a \in A} n(a, X_\tau) \leqslant \frac{\ell}{L}(m - \varepsilon k) \right]. \tag{10}$$

Доказательство. Заметим, что все алгоритмы из $A(X_\tau)$ имеют одинаковое переобучение. Это следует из двух утверждений: во-первых, все алгоритмы из A имеют равное число ошибок на полной выборке, во-вторых, все алгоритмы из $A(X_\tau)$ имеют равное число ошибок на обучении. Значит (9) можно упростить следующим образом:

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \left[\delta(a', X_\tau) \geqslant \varepsilon \right],$$

где a' — произвольный алгоритм из $A(X_\tau)$. Из $a' \in A(X_\tau)$ следует, что $n(a', X_\tau) = \min\limits_{a \in A} n(a,X)$. Подставляя это выражение в определение уклонения частоты $\delta(a,X)$ и проводя элементарные выкладки, получаем (10). ■

Вернемся к выводу формулы вероятности переобучения для центрального слоя хэммингова шара B_r^m с центром в a_0. Обозначим через $X^m = \{x \in \mathbb{X} \colon a_0(x) = 1\}$ множество объектов, на которых ошибается алгоритм a_0, и $X^{L-m} = \{x \in \mathbb{X} \colon a_0(x) = 0\}$ — множество объектов, на которых a_0 не ошибается.

Лемма 10. *Группа $S_m \times S_{L-m}$, где S_m и S_{L_m} — симметрические группы перестановок, действующие на множествах X^m и X^{L-m}, соответственно, является подгруппой группы симметрий множества алгоритмов B_r^m.*

Доказательство. Доказательство следует из определения центрального слоя хэммингова шара и инвариантности хэммингова расстояния к действию группы S_L. ■

Лемма 11. *Орбиты $\tau \in \Omega([\mathbb{X}]^\ell)$ индексированы параметром $i = |X \cap X^m|$. Мощность орбиты τ_i записывается в виде $|\tau_i| = C_L^\ell h_L^{\ell,m}(i)$, где $h_L^{\ell,m}(i) = C_m^t C_{L-m}^{\ell-i}/C_L^\ell$ — функция плотности гипергеометрического распределения.*

Доказательство. Первое утверждение леммы непосредственно следует из структуры подгруппы симметрий, полученной в лемме 10. Мощность орбиты $|\tau_i| = C_L^\ell h_L^{\ell,m}(i)$ определяется числом способов выбрать i объектов из X^m и $\ell - i$ объектов из X^{L-m}. ■

Теперь мы можем приступить к доказательству теоремы 5.

Доказательство. Воспользовавшись теоремой 9 и леммой 11, получим

$$Q_\varepsilon(B_r^m) = \sum_{i=0}^{m} h_L^{\ell,m}(i) \left[\min_{a\in A} n(a, X_i) \leqslant \frac{\ell}{L}(m - \varepsilon k)\right].$$

Напомним, что по определению параметр i обозначает мощность множества $X \cap X^m$. Пусть $r' = \left\lfloor \frac{r}{2} \right\rfloor$. Тогда

$$\min_{a\in B_r^m} n(a, X_i) = \begin{cases} 0, \text{ при } i \leqslant r', \\ i - r', \text{ при } i > r'. \end{cases}$$

Следовательно,

$$Q_\varepsilon(B_r^m) = \begin{cases} 0, \text{ при } m < \varepsilon k, \\ H_L^{\ell,m}(\frac{\ell}{L}(m - \varepsilon k) + \left\lfloor r/2 \right\rfloor), \text{ при } m \geqslant \varepsilon k. \end{cases} ■$$

Из доказанной формулы следует, что вероятность переобучения $Q_\varepsilon(B_r^m)$ соответствует FC-оценке (3) для фиксированного алгоритма, смещенной вправо на $\Delta\varepsilon = \frac{L}{k\ell}\left\lfloor r/2 \right\rfloor$. Это объясняет, почему в эксперименте 2 медиана распределения $Q_\varepsilon(B_r^m)$ растет линейно с ростом радиуса r.

Учёт расслоения

Результаты двух предыдущих параграфов позволяют сформулировать следующую оценку вероятности переобучения.

Теорема 12. *Пусть $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$, и в каждом подмножестве A_i алгоритмы допускают равное число ошибок. Пусть каждое множество A_i вложено в центральный слой хэммингова шара $B_{r_i}^{m_i}$. Тогда*

$$Q_\varepsilon(A) \leqslant \sum_{i=1}^{t} Q_\varepsilon(B_{r_i}^{m_i}) = \sum_{i=1}^{t} \left\{ H_L^{\ell,m_i}(\tfrac{\ell}{L}(m_i - \varepsilon k) + \left\lfloor r_i/2 \right\rfloor) \cdot [m_i \geqslant \varepsilon k] \right\}. \tag{11}$$

Доказательство. Доказательство следует из лемм 3, 4 и формулы (7). ■

Оценка (11) учитывает сходство, но не расслоение алгоритмов по числу ошибок. Этот недостаток можно исправить, применив метод порождающих и запрещающих множеств [5]. Для этого метод необходимо обобщить, чтобы он был применим не к отдельным алгоритмам, а непосредственно к кластерам алгоритмов.

Гипотеза 1. Пусть множество алгоритмов A представлено в виде разбиения на непересекающиеся подмножества $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$. Пусть выборка $\mathbb{X}$ и метод обучения μ

таковы, что для каждого $i = 1, \ldots, t$ можно указать пару непересекающихся подмножеств $X_i \subset \mathbb{X}$ и $X'_i \subset \mathbb{X}$, удовлетворяющую условию

$$[\mu(A, X) \in A_i] \leqslant [X_i \subset X][X'_i \subset \bar{X}], \quad \forall X \in [\mathbb{X}]^\ell.$$

Пусть, кроме этого, все алгоритмы $a \in A_i$ не допускают ошибок на X_i и ошибаются на всех объектах из X'_i.

Множество X_i будем называть *порождающим*, множество X'_i — *запрещающим* для A_i. Гипотеза 1 означает, что результат обучения может принадлежать A_i только в том случае, если в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{Y}_i \equiv \mathbb{X} \backslash X_i \backslash X'_i$ будем называть *нейтральными* для A_i.

Пусть $L_i = L - |X_i| - |X'_i|$, $\ell_i = \ell - |X_i|$, $k_i = k - |X'_i|$. Пусть $Q'_\varepsilon(A_i)$ есть вероятность переобучения на множестве нейтральных объектов $\mathbb{Y}_i$:

$$Q'_\varepsilon(A_i) = \frac{1}{C^{\ell_i}_{L_i}} \sum_{Y \in [\mathbb{Y}_i]^{\ell_i}} [\delta(\mu(A_i, Y), Y) \geqslant \varepsilon],$$

где $[\mathbb{Y}_i]^{\ell_i}$ — множество разбиений $\mathbb{Y}_i$ на обучающую выборку Y длины ℓ_i и контрольную выборку $\bar{Y}$ длины $k_i = L_i - \ell_i$.

Теорема 13 (Оценка расслоения-сходства). *Пусть выполнена гипотеза 1, а на разбиение $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ наложено дополнительное ограничение: внутри каждого кластера A_i все алгоритмы допускают равное число ошибок (обозначаемое через m_i). Тогда вероятность переобучения $Q_\varepsilon(A)$ ограничена сверху следующей оценкой:*

$$Q_\varepsilon(A) \leqslant \sum_{i=1}^{t} P_i \, Q'_{\varepsilon_i}(A_i), \tag{12}$$

где $P_i = \frac{C^{\ell_i}_{L_i}}{C^{\ell}_{L}}$, $\varepsilon_i = \frac{L_i}{\ell_i k_i} \frac{\ell k}{L} \varepsilon + \left(1 - \frac{\ell L_i}{L \ell_i}\right) \frac{m_i}{k_i} - \frac{|X'_i|}{k_i}$, $Q'_\varepsilon(A_i)$ — определенная выше вероятность переобучения на множестве нейтральных объектов.

Доказательство. Распишем определение вероятности переобучения:

$$\begin{aligned} Q_\varepsilon(A) &= \frac{1}{C^{\ell}_{L}} \sum_{X \in [\mathbb{X}]^\ell} [\delta(\mu(A, X), X) \geqslant \varepsilon] = \\ &= \frac{1}{C^{\ell}_{L}} \sum_{i=1}^{t} \sum_{X \in [\mathbb{X}]^\ell} [\mu(A, X) \in A_i][\delta(\mu(A_i, X), X) \geqslant \varepsilon] \leqslant \\ &\leqslant \frac{1}{C^{\ell}_{L}} \sum_{i=1}^{t} \sum_{X \in [\mathbb{X}]^\ell} [X_i \subset X][X'_i \subset \bar{X}][\delta(\mu(A_i, X), X) \geqslant \varepsilon]. \end{aligned}$$

Пусть $Y = X \backslash X_i$. Тогда $\sum\limits_{X \in [\mathbb{X}]^\ell}$ при условии $[X_i \subset X][X'_i \subset \bar{X}]$ можно заменить на суммирование по $Y \in [\mathbb{Y}_i]^{\ell_i}$.

$$Q_\varepsilon(A) \leqslant \frac{C^{\ell_i}_{L_i}}{C^{\ell}_{L}} \sum_{i=1}^{t} \frac{1}{C^{\ell_i}_{L_i}} \sum_{Y \in [\mathbb{Y}_i]^{\ell_i}} [\delta(\mu(A_i, X), X) \geqslant \varepsilon], \text{ где } X = Y \sqcup X_i. \tag{13}$$

Выразим условие $\delta(\mu(A_i, X), X) \geqslant \varepsilon$ в терминах Y. Обозначим $a = \mu(A_i, X)$, и пусть $n(a, Y) = s$. Тогда, используя условие $n(a, X_i) = 0$ и $n(a, X'_i) = |X'_i|$ из гипотезы 1, получим $n(a, X) = s$, $n(a, \bar{X}) = m_i - s$, $n(a, \bar{Y}) = m_i - |X'_i| - s$. Следовательно, условия переобучения для X и Y запишутся следующим образом:

$$[\delta(\mu(A_i, X), X) \geqslant \varepsilon] = \Big[s \leqslant \frac{\ell}{L}(m_i - \varepsilon k)\Big],$$
$$[\delta(\mu(A_i, Y), Y) \geqslant \varepsilon_i] = \Big[s \leqslant \frac{\ell_i}{L_i}\big(m_i - |X'_i| - \varepsilon_i k_i\big)\Big].$$

Пусть $\varepsilon_i = \frac{L_i}{\ell_i k_i}\frac{\ell k}{L}\varepsilon + \big(1 - \frac{\ell L_i}{L\ell_i}\big)\frac{m_i}{k_i} - \frac{|X'_i|}{k_i}$. Непосредственной проверкой убеждаемся, что $[\delta(\mu(A_i, X), X) \geqslant \varepsilon] = [\delta(\mu(A_i, Y), Y) \geqslant \varepsilon_i]$. Подставляя это в (13), получаем утверждение теоремы. ■

Покажем, как для произвольного разбиения $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ построить систему порождающих и запрещающих множеств. Следуя [5], введем на A отношение частичного порядка: $a \leqslant b$ тогда и только тогда, когда $I(a, x) \leqslant I(b, x), \forall x \in \mathbb{X}$. Определим $a < b$ если $a \leqslant b$ и $a \neq b$. Если $a < b$ и при этом $\rho(a, b) = 1$, то будем говорить, что a *предшествует* b, и записывать $a \prec b$.

Для отдельного алгоритма $a \in A$ порождающие и запрещающие множества определены в [5]:

$$\begin{aligned} X_a &= \{x \in X \colon \exists b \in A \colon a \prec b, I(a, x) < I(b, x)\}, \\ X'_a &= \{x \in X \colon \exists b \in A \colon b < a, I(b, x) < I(a, x)\}. \end{aligned} \tag{14}$$

Для кластера A_i положим

$$X_i = \bigcap_{a \in A_i} X_a, \quad X'_i = \bigcap_{a \in A_i} X'_a. \tag{15}$$

Лемма 14. *Множества X_i и X'_i, определенные в* (15)*, являются, соответственно, порождающим и запрещающим множествами для кластера A_i в смысле гипотезы 1.*

Доказательство. Для произвольного разбиения $X \in [\mathbb{X}]^\ell$ обозначим $a = \mu X$, и пусть $a \in A_i$. В [5] показано, что определенные в (14) множества X_a и X'_a являются порождающим и запрещающим множествами для алгоритма a, т.е. из условия $\mu X = a$ следует, что $X_a \subset X$ и $X'_a \subset \bar{X}$. Из определения X_i и X'_i следует, что $X_i \subset X_a$ и $X'_i \subset X'_a$. Следовательно, $X_i \subset X$ и $X'_i \subset \bar{X}$.

Условие «все алгоритмы $a \in A_i$ не допускают ошибок на X_i и ошибаются на всех объектах из X'_i» также следует из определения X_i и X'_i. ■

Полученные выше результаты позволяют уточнить оценку (11).

Теорема 15. *В условиях теоремы 12, определений* (14) *и* (15) *для порождающих и запрещающих множеств справедлива следующая оценка вероятности переобучения:*

$$Q_\varepsilon(A) \leqslant \sum_{i=1}^{t} \frac{C_{L_i}^{\ell_i}}{C_L^\ell} Q'_{\varepsilon'}(B_{r_i}^{m'_i}) = \sum_{i=1}^{t} \Big\{ \frac{C_{L_i}^{\ell_i}}{C_L^\ell} H_{L_i}^{\ell_i, m'_i}(s(\varepsilon) + \lfloor r_i/2 \rfloor) \cdot [m_i \geqslant \varepsilon k] \Big\}, \tag{16}$$

где $s(\varepsilon) = \frac{\ell}{L}(m_i - \varepsilon k)$, $L_i = L - |X_i| - |X'_i|$, $\ell_i = \ell - |X_i|$, $m'_i = m_i - |X'_i|$.

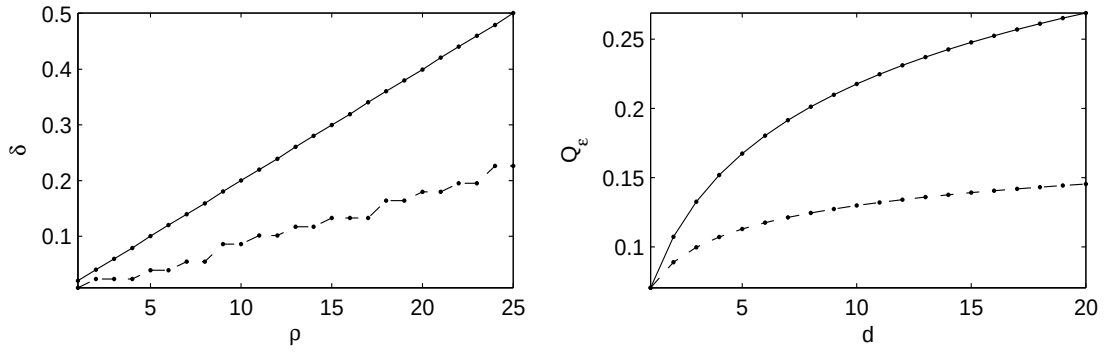

Рис. 2. Переобучение центрального слоя шара $B^m_{2\rho}$ (сплошная кривая) и локальной окрестности $\hat{B}^{m-\rho}_{2\rho,\rho}$ (пунктирная кривая) при $L = 200$, $\ell = 100$, $m = 50$. Рисунок слева отображает среднее уклонение частот ошибок на обучении и контроле в зависимости от параметра ρ. Рисунок справа отображает зависимость средней вероятности переобучения $\bar{Q}_\varepsilon(B, d)$ от параметра d при $\rho = 5$, $\varepsilon = 0.1$.

Доказательство. Доказательство следует из теоремы 13, леммы 14 и формулы (7). ■

Главное отличие (16) от полученной ранее оценки (11) — в коэффициенте $\frac{C^{\ell_i}_{L_i}}{C^{\ell}_{L}}$, экспоненциально убывающем с ростом мощности порождающего и запрещающего множеств.

Локальная окрестность малой мощности

Отметим, что центральный слой хэммингова шара B^m_r даёт достаточно грубую оценку вероятности переобучения для множества $A_i \subset B^m_r$. Во-первых, такая аппроксимация множества A_i не учитывает объекты выборки, лежащие глубоко внутри своего класса, и потому одинаково классифицируемые всеми алгоритмами кластера A_i. Во-вторых, при оценке не учитывается мощность кластера A_i, которая на реальных данных оказывается много меньше мощности множества B^m_r.

Определение 3. *Пусть все объекты генеральной выборки $\mathbb{X}$ разделены на три непересекающихся множества: надежно классифицируемые объекты X_0, ошибочно классифицируемые объекты X_1 и пограничные объекты X_r. Пусть $|X_r| = r$ и $|X_1| = m$, ρ — целочисленный параметр, $\rho \leqslant r$. Рассмотрим алгоритм a_0, допускающий m ошибок на X_1 и ρ ошибок на X_r. Локальной окрестностью алгоритма a_0 будем называть множество алгоритмов $\hat{B}^m_{r,\rho} \subset \mathbb{A}$, удовлетворяющее следующим условиям:*

- *$\hat{B}^m_{r,\rho}$ содержит все алгоритмы, допускающие ровно ρ ошибок на объектах из X_r,*
- *ни один алгоритм из $\hat{B}^m_{r,\rho}$ не ошибается на объектах из X_0,*
- *все алгоритмы из $\hat{B}^m_{r,\rho}$ ошибаются на всех объектах из X_1.*

На рис. 2 слева сравниваются вероятности переобучения центрального слоя хэммингова шара и локальной окрестности. Видно, что локальная окрестность даёт меньшую оценку вероятности переобучения. Следовательно, аппроксимация кластеров A_i с помощью локальных окрестностей даёт более точную оценку вероятности переобучения.

Для дальнейшего уточнения оценки мы будем рассматривать кластер A_i как случайное подмножество некоторого объемлющего множества B (например, центрального слоя хэммингова шара или локальной окрестности).

Определение 4. *Средней вероятностью переобучения по подмножествам фиксированной мощности назовём следующую величину:*

$$\bar{Q}_\varepsilon(B,d) = \frac{1}{C^d_{|B|}} \sum_{A' \in [B]^d} Q_\varepsilon(A'), \tag{17}$$

где $[B]^d = \{A' \subset B \colon |A'| = d\}$ *— система подмножеств фиксированной мощности.*

На рис. 2 справа показан пример зависимости средней вероятности переобучения $\bar{Q}_\varepsilon(B,d)$ от параметра d для двух рассматриваемых нами семейств алгоритмов.

Теорема 16. *Пусть* B *— множество алгоритмов, допускающих равное число ошибок на полной выборке. Тогда выполнено следующее:*

$$\bar{Q}_\varepsilon(B,d) = 1 - \frac{1}{C^\ell_L} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| \frac{C^d_{N_\varepsilon(B,X_\tau)}}{C^d_{|B|}}, \tag{18}$$

где $\Omega([\mathbb{X}]^\ell)$ *— множество орбит действия группы симметрий* A *на* $[\mathbb{X}]^\ell$, X_τ *— произвольный представитель орбиты* $\tau \in \Omega([\mathbb{X}]^\ell)$, $N_\varepsilon(B,X) = \sum\limits_{a \in B} [n(a,X) > s(\varepsilon)]$ *— число алгоритмов из* B*, непереобученных на разбиении* X, $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$.

Доказательство. Запишем определение $\bar{Q}_\varepsilon(B,d)$ и воспользуемся тем, что $A' \subset B$ вновь является подмножеством слоя:

$$\bar{Q}_\varepsilon(B,d) = \frac{1}{C^d_{|B|}} \sum_{A' \in [B]^d} \frac{1}{C^\ell_L} \sum_{X \in [\mathbb{X}]^\ell} \left[\exists a \in A' \colon n(a,X) \leqslant s(\varepsilon)\right].$$

Переставим местами знаки суммирования и применим логическое отрицание:

$$\bar{Q}_\varepsilon(B,d) = 1 - \frac{1}{C^\ell_L} \frac{1}{C^d_{|B|}} \sum_{X \in [\mathbb{X}]^\ell} \underbrace{\sum_{A' \in [B]^d} \left[\forall a \in A' \colon n(a,X) > s(\varepsilon)\right]}_{F_\varepsilon(B,X)}.$$

Заметим, что выделенное в прошлой формуле выражение $F_\varepsilon(B,X)$ соответствует числу способов выбрать из B подмножество A' мощности d так, чтобы ни один из алгоритмов в A' не был переобученным. Обозначим через $N_\varepsilon(B,X)$ общее число алгоритмов в B, непереобученных на разбиении X. Тогда $F_\varepsilon(B,X)$ является числом сочетаний из $N_\varepsilon(B,X)$ по d:

$$\bar{Q}_\varepsilon(B,d) = 1 - \frac{1}{C^\ell_L} \frac{1}{C^d_{|B|}} \sum_{X \in [\mathbb{X}]^\ell} C^d_{N_\varepsilon(B,X)}.$$

По теоремам 7 и 8 функция $C^d_{N_\varepsilon(B,X)}$, где $N_\varepsilon(B,X) = \sum\limits_{a \in B} [n(a,X) > \varepsilon]$, является симметричной функцией третьего рода, и, следовательно, (18) факторизуется по действию группы симметрий на множестве разбиений:

$$\bar{Q}_\varepsilon(B,d) = 1 - \frac{1}{C^\ell_L} \frac{1}{C^d_{|B|}} \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} |\tau| C^d_{N_\varepsilon(B,X_\tau)}.$$

■

При больших значениях параметра d дробь $\frac{C^d_{N_\varepsilon(B,X)}}{C^d_{|B|}}$ приближенно равна $\left(\frac{N_\varepsilon(B,X_\tau)}{|B|}\right)^d$. Таким образом, вклад разбиения $(X,\bar{X})$ в вероятность переобучения полностью определяется мощностью d рассматриваемых подмножеств и числом алгоритмов из B, непереобученных на разбиении $(X,\bar{X})$.

Покажем, как вычислять оценку (18) на примере локальной окрестности $\hat{B}^m_{r,\rho}$.

Теорема 17. *Средняя вероятность переобучения случайного подмножества $A_i \subset \hat{B}^m_{r,\rho}$ фиксированной мощности d даётся следующей формулой:*

$$\bar{Q}_\varepsilon(\hat{B}^m_{r,\rho},d) = 1 - \frac{1}{C^\ell_L}\sum_{i=0}^{\min(m,\ell)}\sum_{j=0}^{\min(r,\ell-i)} C^i_m C^j_r C^{\ell-i-j}_{L-m-r}\frac{C^d_{N_{i,j}}}{C^d_{|\hat{B}^m_{r,\rho}|}}, \tag{19}$$

где

$$N_{i,j} = \sum_{t=0}^{\min(j,\rho)} C^t_j C^{\rho-t}_{r-j}[i+t > s(\varepsilon)].$$

Доказательство. Рассмотрим три симметрические группы перестановок S_m, S_r и S_{L-m-r}, действующие на множествах X_1, X_r и X_0, соответственно. Группой симметрий множества алгоритмов $\hat{B}^m_{r,\rho}$ является декартово произведение $S_m \times S_r \times S_{L-m-r}$. Орбиты действия $\mathrm{Sym}(\hat{B}^m_{r,\rho})$ на $[\mathbb{X}]^\ell$ индексируются двумя параметрами, $i = |X \cap X_1|$ и $j = |X \cap X_r|$, где X — обучающая выборка. Мощность орбиты $\tau_{i,j}$ даётся, соответственно, выражением $|\tau_{i,j}| = C^i_m C^j_r C^{\ell-i-j}_{L-m-r}$.

Разобравшись с симметриями, необходимо для представителя $X_{i,j} \in \tau_{i,j}$ вычислить величину $N_{i,j} = N(\hat{B}^m_{r,\rho}, X_{i,j})$ — количество алгоритмов из $\hat{B}^m_{r,\rho}$, непереобученных на разбиении $X_{i,j}$. Рассмотрим произвольный алгоритм $a \in \hat{B}^m_{r,\rho}$ и обозначим через t количество ошибок данного алгоритма на $X \cap X_r$. Тогда данный алгоритм делает $i+t$ ошибок на обучении, и, следовательно, условие того, что он не переобучен, записывается в виде $i+t > s(\varepsilon)$, где $s(\varepsilon) = \frac{\ell}{L}(m+\rho-\varepsilon k)$. Количество алгоритмов в $\hat{B}^m_{r,\rho}$ с данным значением параметра t равно $C^t_j C^{\rho-t}_{r-j}$. Суммируя по t, получим количество непереобученных алгоритмов:

$$N_{i,j} = \sum_{t=0}^{\min(j,\rho)} C^t_j C^{\rho-t}_{r-j}[i+t > s(\varepsilon)].$$

Тогда, по теореме 16, вероятность переобучения случайного подмножества $\hat{B}^m_{r,\rho}$, состоящего из d алгоритмов, даётся следующей формулой:

$$\bar{Q}_\varepsilon(\hat{B}^m_{r,\rho},d) = 1 - \frac{1}{C^\ell_L}\sum_{i=0}^{\min(m,\ell)}\sum_{j=0}^{\min(r,\ell-i)} |\tau_{i,j}|\frac{C^d_{N_{i,j}}}{C^d_{|\hat{B}^m_{r,\rho}|}}, \tag{20}$$

где $\tau_{i,j}$ и $N_{i,j}$ определены выше. ■

Теорема 17 позволяет уточнить верхнюю оценку вероятности переобучения (16). Для этого будем оценивать $Q_\varepsilon(A_i)$ с помощью $\bar{Q}_\varepsilon(B^m_{r,\rho})$, где d — мощность A_i, m — число объектов, на которых ошибаются все $a \in A_i$, r — число таких объектов, для которых хотя бы два алгоритма из A_i дают разную классификацию, ρ — среднее число ошибок алгоритмов из A_i на множестве X_r. Отметим, что в данном случае оценка $Q_\varepsilon(A_i) \leqslant \bar{Q}_\varepsilon(\hat{B}^m_{r,\rho})$ является эвристикой.

Эксперимент

Проведем экспериментальное сравнение оценки расслоения-сходства (12) с тремя комбинаторными оценками (VC- и SC-оценками из [6], ES-оценкой из [14]) и двумя PAC-Bayesian оценками из [15].

В качестве исходных данных были взяты 11 задач из репозитория UCI [16]. Описание задач приводится в таблице 2. На этапе предобработки удалялись объекты с хотя бы одним пропущенным признаком. После этого каждый признак нормировался в интервал $[0, 1]$. Для многоклассовых задач целевые классы были вручную сгруппированы в два класса.

Таблица 2. Описание задач.

Задача	#Объектов	#Признаков	Задача	#Объектов	#Признаков
Glass	214	9	Statlog	2310	19
Liver dis.	345	6	Wine	4898	11
Ionosphere	351	34	Waveform	5000	21
Australian	690	6	Pageblocks	5473	10
Pima	768	8	Optdigits	5620	64
Faults	1941	27			

Для каждой задачи мы выполняли процедуру пятикратной кросс-валидации, которая запускалась 100 раз для усреднения результатов. Таким образом, для каждой задачи мы генерировали $M = 500$ разбиений $\mathbb{X} = \mathbb{X}^i_L \sqcup \mathbb{X}^i_K$, $i = 1, \ldots, M$ и вычисляли оценку Монте-Карло для среднего уклонения частот ошибок логистической регрессии:

$$\hat{\delta}_L(\mu_{\text{LR}}, \mathbb{X}) = \frac{1}{M} \sum_{i=1}^{M} \nu(\mu_{\text{LR}} \mathbb{X}^i_L, \mathbb{X}^i_K) - \nu(\mu_{\text{LR}} \mathbb{X}^i_L, \mathbb{X}^i_L).$$

После этого каждая обучающая выборка $\mathbb{X}_L$ использовалась для вычисления комбинаторной оценки на уклонение частот ошибок МЭР. Множества алгоритмов A, из которого МЭР выбирал лучший алгоритм, генерировалось с помощью случайных блужданий по графу расслоения-связности линейных классификаторов [14]. В качестве начального приближения для случайного блуждания использовался алгоритм $\mu_{\text{LR}}\mathbb{X}_L$, настроенный логистической регрессией по обучающей выборке $\mathbb{X}_L$. Далее вновь использовался метод Монте-Карло: генерировались $M' = 4096$ случайных разбиений $\mathbb{X}_L = X^j_\ell \sqcup X^j_k$, $j = 1, \ldots, M'$ (при $\frac{\ell}{L} = 0.8$) и вычислялась следующая величина:

$$\hat{\delta}_\ell(\mu, \mathbb{X}_L) = \frac{1}{M'} \sum_{j=1}^{M'} \nu(\mu X^j_\ell, X^j_k) - \nu(\mu X^j_\ell, X^j_\ell),$$

где μ — метод минимизации эмпирического риска. В заключение эта величина усреднялась по всем разбиениям $\mathbb{X} = \mathbb{X}^i_L \sqcup \mathbb{X}^i_K$. Величины $\hat{\delta}_L(\mu_{\text{LR}}, \mathbb{X})$ и $\hat{\delta}_\ell(\mu, \mathbb{X}_L)$ соответствуют реальному переобучению логистической регрессии и его идеальной оценке в рамках комбинаторного подхода.

В таблице 3 сравниваются следующие величины: $\hat{\delta}_L(\mu_{\text{LR}}, \mathbb{X})$ и $\hat{\delta}_\ell(\mu, \mathbb{X}_L)$; четыре верхние комбинаторные оценки величины $\hat{\delta}_\ell(\mu, \mathbb{X}_L)$, обозначенные как VC, SC, ES и AF; две PAC-Bayes оценки (обозначены как DD и DI). В отличие от комбинаторных оценок, ограничивающих уклонение частот ошибок, PAC-Bayes оценки справедливы непосредственно для частоты ошибок на контроле (приведена в столбце «Ошибка тест»). DD-оценка

Таблица 3. Сравнение фактического переобучения и различных оценок.

	Ошибка	Переобучение		Комбинаторные оценки				PAC-Bayes	
Task	Тест	$\hat{\delta}_L(LR)$	$\hat{\delta}_\ell(\mu)$	VC	SC	ES	AF	DI	DD
glass	0.076	0.030	0.067	0.191	0.127	0.124	0.106	1.268	0.740
Liver dis.	0.315	0.017	0.046	0.249	0.192	0.146	0.161	1.207	1.067
Ionosphere	0.126	0.079	0.042	0.138	0.099	0.087	0.084	1.219	1.149
Australian	0.136	0.014	0.023	0.130	0.101	0.081	0.086	1.145	0.678
pima	0.227	0.007	0.021	0.151	0.117	0.090	0.098	0.971	0.749
faults	0.210	0.011	0.008	0.091	0.070	0.046	0.060	1.110	1.054
statlog	0.142	0.004	0.008	0.072	0.060	0.043	0.051	1.102	0.746
wine	0.250	0.002	0.003	0.061	0.047	0.032	0.040	0.776	0.637
waveform	0.105	0.003	0.003	0.043	0.033	0.023	0.023	0.561	0.354
pageblocks	0.051	0.001	0.003	0.030	0.022	0.016	0.018	0.739	0.186
Optdigits	0.121	0.006	0.003	0.043	0.034	0.023	0.026	1.068	0.604

учитывает размерность пространства признаков и является более точной по сравнению с универсальной DI-оценкой, справедливой для любого числа признаков.

Комбинаторная SC-оценка соответствует оценке расслоения-связности из [6]. VC-оценка также приведена в [6], и она, в отличие от SC-оценки, не учитывает ни расслоение, ни связность. ES-оценка [14] основана на более тонком учёте расслоения, при котором каждый алгоритм сравнивается со всем множеством найденных истоком графа расслоения-связности.

AF-оценка, предлагаемая нами в данной статье, получена из оценки расслоения-сходства (12). Чтобы полностью конкретизировать оценку (12), необходимо уточнить следующее: метод разбиения исходного множества алгоритмов A на кластеры, способ выбора порождающих и запрещающих множеств для каждого кластера, способ оценивания вероятности переобучения каждого кластера. В AF-оценке порождающие и запрещающие множества выбирались в соответствии с (15), для оценки вероятности переобучения каждого кластера используется формула (19), а представление множества алгоритмов A в виде $A = A_1 \sqcup \cdots \sqcup A_t$ производится с помощью иерархической кластеризации при выборе расстояния дальнего соседа. Исходная метрика на A определяется как хэммингово расстояние между векторами ошибок алгоритмов.

Из экспериментальных данных следует, что переобучение логистической регрессии хорошо приближается комбинаторной оценкой $\hat{\delta}_\ell(\mu, \mathbb{X}_L)$ для МЭР. Все четыре комбинаторные оценки существенно точнее обеих PAC-bayes оценок. Среди комбинаторных оценок наименее завышенной оказывается ES-оценка. За ней следует предложенная нами AF-оценка. Каждая из этих оценок существено уточняет SC-оценку расслоения-связности. Улучшение точности в ES- и AF-оценках основано на двух различных эффектах — более тонком учете расслоения в ES-оценке и учете сходства алгоритмов в AF-оценке. Представляется возможным, что объединение ES- и AF-оценок позволит добиться еще большего качества комбинаторных оценок вероятности переобучения.

Выводы

В данной работе предложена новая оценка вероятности переобучения, учитывающая расстояния между векторами ошибок алгоритмов. Оценка основана на разложении множества алгоритмов на кластеры, т.е. на непересекающиеся подмножества, состоящие из алго-

ритмов с похожими векторами ошибок. Для каждого такого кластера применяется верхняя оценка вероятности переобучения, учитывающая хэммингов диаметр кластера и его мощность. По аналогии с уже существующими оценками расслоения-связности доказана новая оценка, одновременно учитывающая и эффект расслоения по числу ошибок, и эффект сходства алгоритмов. Вывод данной оценки существенно опирается на теоретико-групповой подход. Эффективность полученных оценок продемонстрирована на примере 11 задач из репозитория UCI. Показано, что предлагаемый метод в ряде случаев даёт более точную оценку переобучения по сравнению с уже известным оценками.

Литература

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[2] *Vapnik V.* Statistical Learning Theory. — New York: Wiley, 1998.

[3] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[4] *Vorontsov K. V.* Splitting and Similarity Phenomena in the Sets of Classifiers and Their Effect on the Probability of Overfitting // Pattern Recognition and Image Analysis. — 2009. — Vol. 19, No. 3. — Pp. 412–420.

[5] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 269–285.

[6] *Vorontsov K. V., Ivahnenko A. A., Reshetnyak I. M.* Generalization bound based on the splitting and connectivity graph of the set of classifiers // Pattern Recognition and Image Analysis: new information technologies (PRIA-10), St. Petersburg, Russian Federation, December 5–12, 2010.

[7] *Фрей А. И.* Вероятность переобучения плотных и разреженных многомерных сеток алгоритмов // Междунар. конф. ИОИ-8 — М.: МАКС Пресс, 2010. — С. 87–90.

[8] *Vorontsov K. V., Ivahnenko A. A.* Tight Combinatorial Generalization Bounds for Threshold Conjunction Rules // 4-th International Conference on Pattern Recognition and Machine Intelligence, June 27 – July 1, 2011. Lecture Notes in Computer Science. Springer-Verlag, 2011. — Pp. 66–73.

[9] *Фрей А. И., Ивахненко А. А, Решетняк И. М.* Применение комбинаторных оценок вероятности переобучения в простом голосовании конъюнкций // Междунар. конф. ИОИ-9 — М.: МАКС Пресс, 2012. — С. 86-89.

[10] *E. Bax.* Similar Classifiers and VC Error Bounds // Tech. Rep. CalTech-CS-TR97-14: 1997.

[11] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 66–69.

[12] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированных методов обучения // Pattern Recognition and Image Analysis. — 2010.

[13] *Толстихин И. О.* Вероятность переобучения плотных и разреженных семейств алгоритмов // Междунар. конф. ИОИ-8 — М.: МАКС Пресс, 2010. — С. 83–86.

[14] *Соколов Е. А., Воронцов К. В.* Минимизация вероятности переобучения для композиций линейных классификаторов малой размерности // Междунар. конф. ИОИ-9 — М.: Торус Пресс, 2012. — С. 82–85.

[15] Jin C., Wang L. (2012) Dimensionality Dependent PAC-Bayes Margin Bound. *In Advances in Neural Information Processing Systems*, 25, 1043–1051.

[16] *K. Bache, M. Lichman.* UCI Machine Learning Repository // University of California, Irvine, School of Information and Computer Sciences. — 2013.